IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 25, 2021

W. Wang
A. Wang
China Telecom
H. Wang
Huawei Technologies
G. Mishra
Verizon Inc.
S. Zhuang
J. Dong
Huawei Technologies
August 24, 2020

Route Distinguisher Outbound Route Filter (RD-ORF) for BGP-4
draft-wang-idr-rd-orf-03

Abstract

This draft defines a new Outbound Route Filter (ORF) type, called the Route Distinguisher ORF (RD-ORF). RD-ORF is applicable when the routers do not exchange VPN routing information directly (e.g. routers in single-domain connect via Route Reflector, or routers in Option B/Option AB/Option C cross-domain scenario).

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

With the rapid growth of network scale, Route Reflector is introduced in order to reduce the network complexity. Routers in the same Autonomous System only need to establish iBGP session with RR to transmit routes.

In VPN scenario shown in Figure 1, PE1 - PE4 establish IBGP sessions with RR to ensure the routes can be transmitted within AS100, where PE1 and PE3 maintain VRFs of VPN1 and VPN2, PE2 maintains VPN1's VRF and PE4 maintains VPN2's VRF. RR don not maintain any VRFs.

```
              +----------------------------------------------+
              |                                              |
              |                                              |
              |   +---------+                  +---------+   |
              |   |   PE1   |                  |   PE4   |   |
              |   +---------+                  +---------+   |
              |      VPN1    \                /   VPN2       |
              |      VPN2      \+---------+ /                |
              |                 |         |                  |
              |                 |   RR    |                  |
              |                 |         |                  |
              |                 +---------+ \                |
              |                /              \              |
              |   +---------+/                 +---------+   |
              |   |   PE2   |                  |   PE3   |   |
              |   +---------+                  +---------+   |
              |      VPN1                         VPN1       |
              |                    AS 100         VPN2       |
              +----------------------------------------------+
```

Figure 1: Single-domain scenario

When the VRF of VPN1 in PE1 overflows, due to PE1 and other PEs are not iBGP neighbors, BGP Maximum Prefix Features cannot work, so the problem on PE2 cannot be known.

Now, there are several solutions can be used to alleviate this problem:

- Route Target Constraint (RTC) as defined in [RFC4684]

- Address Prefix ORF as defined in [RFC5292]

- PE-CE edge peer Maximum Prefix

- Configure the Maximum Prefix for each VRF on edge nodes

However, there are limitations to existing solutions:

1) Route Target Constraint

RTC can only filter the VPN routes from the uninterested VRFs, if the "trashing routes" come from the interested VRF, filter on RTs will erase all prefixes from this VRF.

2) Address Prefix ORF

Using Address Prefix ORF to filter VPN routes need to pre-configuration, but it is impossible to know which prefix may cause overflow in advance.

3) PE-CE edge peer Maximum Prefix

This mechanism can only protect the edge between PE-CE, it can't be deployed within PE that peered via RR. Depending solely on the edge protection is dangerous, because if only one of the edge points being comprised/error-configured/attacked, then all of PEs within domain are under risk.

4) Configure the Maximum Prefix for each VRF on edge nodes

When a VRF overflows, PE will break down the BGP session with RR according to the Maximum Prefix mechanism. However, there may have several VRFs on PE rely on the PE-RR session, this mechanism will influence other VRFs.

This draft defines a new ORF-type, called the Route Distinguisher ORF (RD-ORF). Using RD-ORF mechanism, VPN routes of a VPN can be controlled based on source RD and originator. This mechanism is event-driven and does not need to be pre-configured. When a VRF of a router overflows, the router will find out the main source address and RD of VPN routes in this VRF, and send a RD-ORF to its BGP peer that carrys the RD and the source address. If a BGP speaker receives a RD-ORF from its BGP peer, it will filter the VPN routes it tends to send according to the RD-ORF entry.

RD-ORF is applicable when the routers do not exchange VPN routing information directly (e.g. routers in single-domain connect via Route Reflector, or routers in Option B/Option AB/Option C cross-domain scenario).

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] .

3. Terminology

The following terms are defined in this draft:

o RD: Route Distinguisher, defined in [RFC4364]

o ORF: Outbound Route Filter, defined in [RFC5291]

- AFI: Address Family Identifier, defined in [RFC4760]

- SAFI: Subsequent Address Family Identifier, defined in [RFC4760]

- EVPN: BGP/MPLS Ethernet VPN, defined in [RFC7432]

- RR: Router Reflector, provides a simple solution to the problem of IBGP full mesh connection in large-scale IBGP implementation.

- VRF: Virtual Routing Forwarding, a virtual routing table based on VPN instance.

4. RD-ORF Encoding

In this draft, we defined a new ORF type called Route Distinguisher Outbound Route Filter (RD-ORF). The ORF entries are carried in the BGP ROUTE-REFRESH message as defined in [RFC5291]. A BGP ROUTE-REFRESH message can carry one or more ORF entries, and MUST be regenerated when it is tended to be sent to other BGP peers. The ROUTE-REFRESH message which carries ORF entries contains the following fields:

- AFI (2 octets)

- SAFI (1 octet)

- When-to-refresh (1 octet): the value is IMMEDIATE or DEFER

- ORF Type (1 octet)

- Length of ORF entries (2 octets)

A RD-ORF entry contains a common part and type-specific part. The common part is encoded as follows:

- Action (2 bits): the value is ADD, REMOVE or REMOVE-ALL

- Match (1 bit): the value is PERMIT or DENY

- Reserved (5 bits)

RD-ORF also contains type-specific part. The encoding of the type-specific part is shown in Figure 2.

```
+-----------------------------------------+
|                                         |
|          Sequence (4 octets)            |
|                                         |
+-----------------------------------------+
|                                         |
|      Route Distinguisher (8 octets)     |
|                                         |
+-----------------------------------------+
|                                         |
|Source Address sub-TLV (4,6 or 16 octets)|
|                                         |
+-----------------------------------------+
```

Figure 2: RD-ORF type-specific encoding

- Sequence: identifying the order in which RD-ORF is generated

- Route Distinguisher: distinguish the different user routes. The RD-ORF filters the VPN routes it tends to send based on Route Distinguisher.

- Source Address sub-TLV: the source address is TLV format, which contains the following sub-TLVs:

  * In single-domain or Option C cross-domain scenario, NEXT_HOP attribute is fixed during routing transmission, so it can be used as source address.

    Type = 1, Length = 4 or 16 octets, value = NEXT_HOP.

  * In Option B or Option AB cross-domain scenario, NEXT_HOP attribute may be changed by ASBRs and cannot be used as the source address. The originator can be traced by the Route Origin Community in BGP (as defined in Section 5 of [RFC4360]).

    Type = 2, Length = 6 octets, value = the value field of Route Origin Community.

Note that if the Action component of an ORF entry specifies REMOVE-ALL, the ORF entry does not include the type-specific part.

When the BGP ROUTE-REFRESH message carries RD-ORF entries, it must be set as follows:

- The ORF-Type MUST be set to RD-ORF.

- The AFI MUST be set to IPv4, IPv6, or Layer 2 VPN (L2VPN).

- If the AFI is set to IPv4 or IPv6, the SAFI MUST be set to MPLS-labeled VPN address.

- If the AFI is set to L2VPN, the SAFI MUST be set to BGP EVPN.

- The Match field MUST be equal to DENY.

5. Application in single-domain scenario

5.1. Addition of RD-ORF entries

The operation of RD-ORF mechanism on each device is independent, each of them makes a local judgement to determine whether it needs to send RD-ORF to its peers.

In general, every VRF on PE is configured a Maximum Prefix, the trigger of RD-ORF mechanism can be set as the number of VPN routes in VRF reach 80% of the Maximum Prefix. For RR, it doesn't have VRF and the machanism can be triggered by other conditions, such as the RR's memory/CPU utilization reaches 80%.

When the RD-ORF mechanism is triggered, the device must send an alarm information to network operators.

5.1.1. Operation process of RD-ORF mechanism on source PE

In scenario shown in Figure 1, when the VRF of VPN1 in PE1 overflows, PE1 will do analysis and calculation locally to find out the main source of VPN routes in this VRF, assuming it is PE3. Then, PE1 will resolve the host address and corresponding RD of VPN routes from BGP UPDATE message, and generate a BGP ROUTE-REFRESH message contains a RD-ORF entry, and send it to RR. The message contains the following fields:

- AFI is set to IPv4 , IPv6 or L2 VPN

- SAFI is set to "MPLS-labeled VPN address" or "BGP EVPN"

- When-to-refresh is set to IMMEDIATE

- ORF Type is set to RD-ORF

- Length of ORF entries depends on the type of Source Address sub-TLV (21, 23 or 33 octets)

- Action is set to ADD

- Match is set to DENY

- Sequence is set to 1

- Route Distinguisher is set to RD1

- Source Address sub-TLV is set to PE3's host address

It noted that the Sequence can uniquely identifies an RD-ORF entry. All VRFs share the sequence field, and the corresponding sequence of RD-ORF sent by each VRF will be recorded on the device.

5.1.2. Operation process of RD-ORF mechanism on RR

When RR receives the ROUTE-REFRESH message, it checks <AFI/SAFI, ORF-Type, Sequence, Route Distinguisher, Source Address sub-TLV> to find whether it received the latest entry or not. If not, RR will discard the entry; otherwise, RR will add the RD-ORF entry into its Adj-RIB-out.

Before sending a VPN route toward PE1, RR will check its Adj-RIB-out and find there is a filter associated with <RD1, PE3's host address>. Then, RR will stop sending that VPN route to PE1.

If the processing capacity of RR reaches the limit (e.g. RR's memory/CPU utilization reaches 80%), RR will find out the peer that sends the most routing entries to it, assuming it is PE3. Then, RR will generate a BGP ROUTE-REFRESH message contains a RD-ORF entry based on the result of calculation, and send it to PE3.

5.1.3. Operation process of RD-ORF mechanism on target PE

After receiving the ROUTE-REFRESH message that carries a RD-ORF entry, PE3 will check if it receives the latest entry. If not, PE3 will discard it; otherwise, PE3 will add the RD-ORF entry into its Adj-RIB-out.

Before sending a VPN route toward RR, PE3 will check its Adj-RIB-out and find the RD-ORF entry prevent it from sending VPN route which carries RD1 to RR. Then, PE3 will stop sending that VPN route.

The BGP Maximum Prefix Features can be configured to protect PE-CE peering at the edge. Therefore, in general, CEs will not cause the overflow of PEs. If the boundary protection measures fail and cause the overflow, the PE can calculate and find the CEs in corresponding VRF, and break down the associated BGP sessions.

5.2. Withdraw of RD-ORF entries

When the RD-ORF mechanism is triggered, the alarm information will be generated and sent to the network operators. Operators should manually configure the network to resume normal operation. Due to devices can record the RD-ORF entries sent by each VRF, operators can find the entries needs to be withdrawn, and trigger the withdraw process as described in [RFC5291] manually to delete them on RR/ASBR/target PE after network recovery.

6. Applications in cross-domain scenarios

6.1. Application in Option B/Option AB cross-domain scenario

The Option B/Option AB cross-domain scenario is shown in Figure 3:

```
   +--------------------------+             +--------------------------+
   |                          |             |                          |
   |                          |             |                          |
   |   +---------+            |             |            +---------+   |
   |   |   PE1   |            |             |            |   PE3   |   |
   |   +---------+            |             |            +---------+   |
   |      VPN1    \           |             |           /   VPN1       |
   |      VPN2      \+---------+    EBGP   +---------+/     VPN2       |
   |                 |         |           |         |                 |
   |                 |  ASBR1  |-----------|  ASBR2  |                 |
   |                 |         |           |         |                 |
   |                 +---------+           +---------+                 |
   |                /         |             |         \                |
   |   +---------+/           |             |           \+---------+   |
   |   |   PE2   |            |             |            |   PE4   |   |
   |   +---------+            |             |            +---------+   |
   |      VPN1                |             |               VPN2       |
   |           AS1            |             |           AS2            |
   +--------------------------+             +--------------------------+
```

Figure 3: The Option B/Option AB cross-domain scenario

In Option B cross-domain scenario, PE1 - PE4 are responsible for maintaining VPN routing information in AS1 and AS2. There is a direct link between ASBR1 and ASBR2 via EBGP. In AS1, PE1 and PE2 establish IBGP sessions with ASBR1 to ensure the routes can be transmitted in AS1. In AS2, PE3 and PE4 establish IBGP session with ASBR2.

Due to the maintenance of VPN routes is only done by PEs. ASBRs cannot know whether the PEs' ability to handle VPN routes has reached

the upper limit or not, so it needs the RD-ORF to control the number of routes.

Assume that PE1 - PE4 can transmit VPN routes through the network architecture shown in Figure 3. When the VRF of VPN1 in PE1 overflows, the RD-ORF mechanism will be implemented as follows:

1) PE1 will check and find out the main source of VPN routes in this VRF is PE3. Then, PE1 will resolve the host address and corresponding RD from BGP UPDATE message, and generate a BGP ROUTE-REFRESH message contains an RD-ORF entry, and send it to ASBR1.

2) When ASBR1 receives the ROUTE-REFRESH message, it checks whether it receives the latest RD-ORF entry. If not, ASBR1 will discard the entry; Otherwise, ASBR1 will add the RD-ORF entry into its Adj-RIB-out.

Before sending a VPN route toward PE1, RR will check its Adj-RIB-out and find there is a filter associated with <RD1, PE3's host address>. Then, ASBR1 will stop sending that VPN route.

Besides, ASBR1 will locally determine if it needs to send an RD-ORF entry to ASBR2. The judgment criteria refers to Section 5.1.2.

3) If ASBR2/PE3 receives the RD-ORF entry, it will repeat the above process.

When the RD-ORF mechanism is triggered, network operators need to manually configure the network to return to resume normal operation. The withdraw of RD-ORF entries refers to Section 5.2.

In Option AB cross-domain scenario, ASBRs maintain VRFs. However, due to VPN routes in all VRFs use the same BGP session, ASBRs cannot prevent the overflow of a certain VRF by breaking down a BGP session. The operation process of RD-ORF is similar to that in Option B scenario.

6.2. Application in Option C cross-domain scenario

The Option C cross-domain scenario is shown in Figure 4:

```
                                    MP-EBGP
                +----------------------------------------+
                |                                        |
   +------------+------------+              +------------+------------+
   |       +----+----+       |              |       +----+----+       |
   |       |         |       |              |       |         |       |
   |  +----+   RR1   +----+  |              |  +----+   RR2   +----+  |
   |  |    |         |    |  |              |  |    |         |    |  |
   |  |    +---------+    |  |              |  |    +---------+    |  |
   |  |                   |  |              |  |                   |  |
   |  |IBGP           IBGP|  |              |  |IBGP           IBGP|  |
   |  |                   |  |              |  |                   |  |
 +-+--+----+         +----+--+-+          +-+--+----+         +----+--+-+
 |         |         |         |          |         |         |         |
 |   PE1   |         |  ASBR1  |----------|  ASBR2  |         |   PE2   |
 |         |         |         |          |         |         |         |
 +-+-------+   AS1   +-------+-+          +-+-------+   AS2   +-------+-+
   +-------------------------+              +-------------------------+
```

Figure 4: The Option C cross-domain scenario

In this scenario, PE1 and PE2 are responsible for maintaining VPN routing information in AS1 and AS2. In order to reduce the complexity that full-mesh brings to the network, RR1 and RR2 establish MP-EBGP session to transmit labeled routes. In AS1, PE1 and ASBR1 establish IBGP session with RR1 to ensure the routes can be transmitted in AS1. In AS2, PE2 and ASBR2 establish IBGP session with RR2.

Due to the maintenance of VPN routes is only done by PEs. RRs cannot know whether the PEs' ability to handle VPN routes has reached the upper limit or not, so it needs the RD-ORF to control the number of routes.

The operating mechanism of RD-ORF is similar to the description in Section 6.1.

7. Security Considerations

A BGP speaker will maintain the RD-ORF entries in Adj-RIB-out, this behavior consumes its memory and compute resources. To avoid the excessive consumption of resources, [RFC5291] specifies that a BGP speaker can only accept ORF entries transmitted by its interested peers.

8. IANA Considerations

This document defines a new Outbound Route Filter type - Route Distinguisher Outbound Route Filter (RD-ORF). The code point is from the "BGP Outbound Route Filtering (ORF) Types". It is recommended to set the code point of RD-ORF to 66.

IANA is requested to allocate one code point for Source Address sub-TLV for RD-ORF.

This document defines the following new RD-ORF sub-TLV types, which should be reflected in the Source Address sub-TLV for RD-ORF Code Point registry:

Type	Description
1	Next hop Source Address sub-TLV
2	Route Origin Community Source Address sub-TLV

9. Acknowledgement

Thanks Robert Raszuk, Jim Uttaro, Jakob Heitz, Jeff Tantsura, Rajiv Asati, John E Drake and Gert Doering for their valuable comments on this draft.

10. Normative References

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-09 (work in progress), June 2020.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <https://www.rfc-editor.org/info/rfc5291>.

[RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, DOI 10.17487/RFC5292, August 2008, <https://www.rfc-editor.org/info/rfc5292>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

Authors' Addresses

Wei Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangw36@chinatelecom.cn

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: rainsword.wang@huawei.com

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Shunwan Zhuang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: zhuangshunwan@huawei.com

Jie Dong
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: jie.dong@huawei.com